Nutch2Cassandra

Setting up NUTCH 2.x with CASSANDRA

One of the novelties in Nutch 2 is ***Apache Gora*** as a back-end, which provides an in-memory data model and persistence for big data. It allows connecting to different storage options, such as key/value store *Apache Accumulo*, distributed big data store *Apache HBase* and column family data store *Apache Cassandra*. The setting up of Nutch using HBase as a backend is explained in Nutch2Tutorial.

In this tutorial, however, we explain how to run Nutch 2.x using *Cassandra*.

Step 1: Setting up Cassandra

The version used here is: *apache-cassandra-1.2.8-bin.tar.gz*

You can find specific guidance to installation of Cassandra:
here.

Once installed, you should test the installation by starting Cassandra from the konsole using the following command:

(take care to use _ 'sudo'_ unless it was installed without file permission)

- $ cd <install_location>
- $ sudo bin/cassandra (in the background - default)
- $ bin/cassandra -f (in the foreground)

Note: Additionally, to get access to Cassandra tables etc. you can start the *Cassandra Client* by running:

./bin/cassandra-cli -host localhost -port 9160

This should then connect to the _ 'Test Cluster'_ and print the following to the console:

*_ "Connected to: "Test Cluster" on localhost/9160

Welcome to Cassandra CLI version 1.2.8 ..."_*

Further, pressing *?* gives several commandline options, such as:

- *describe cluster;* - shows information on the cluster
- *show keyspaces;* - shows all tables in the cluster

Step 2: Setting up Nutch 2.x

A recent source version of Nutch 2 can be downloaded from here.

It has then to be compiled using *'ant runtime'*.

- Cassandra-specific configuration in Nutch 2.x:
  - In *<Nutch-install>/conf/nutch-site.xml*, specify:

*<property>
<name>storage.data.store.class</name>
<value>org.apache.gora.cassandra.store.CassandraStore</value>
<description>Default class for storing data</description>
</property>*

- Add/uncomment the following properties in *<Nutch-install>/conf/gora.properties* to ensure that Cassandra is set as the default datastore:

*gora.datastore.default=org.apache.gora.cassandra.store.CassandraStore
gora.cassandrastore.servers=localhost:9160*

- Uncomment cassandra-specific entry in *<Nutch-install>/ivy/ivy.xml* to ensure the Cassandra gora-cassandra dependency is available:

  <dependency org="org.apache.gora" name="gora-cassandra" rev="0.3" conf="*->default" />

N.B. run: *'ant runtime'* from the root of the installation folder

---

Crawling in Nutch 2.x

Setting up a basic crawl remains the same as in Nutch 1.x, except that you need to start Cassandra (and the Cassandra client) before starting your crawl.

For instructions for setting up and running a basic crawl: see NutchTutorial (Nutch crawling tutorial with 1.x)

Using the crawl script, crawling can be started from *Nutch-2.x/runtime/deploy/* by running:

bin/crawl <seedDir> <crawlDir> <solrURL> <numberOfRounds>

- where <seedDir> specifies dir + url file, e.g. *urls/seed.txt*, which is your txt file containing the urls you wish to crawl put on hdfs
- <crawlDir>: the folder to query for crawl output ~ crawldb in Nutch 1.x.
- <solrURL>: If you wish to index with solr, otherwise the solr-specific parts have to be removed from the crawl script
- <numberOfRounds>: the number of iterations of generating, fetching and parsing.
  (In 1 iteration, it will process the number specified in the fetchlist, e.g. 50.000 => 50.000 urls per iteration are selected from the crawldb.)

Note: If Nutch 2.x has been successfully running, it should have created a keyspace, called ‘webpage’, which can be viewed in the Cassandra client, when using the command from above: *show keyspaces;*

N.B: If you want to start from scratch, making sure no old urls are re-read from the table, one can remove a table from Cassandra through the client E.g. deleting the table: ‘webpage’ by running: *drop keyspace webpage;*

Checking the results of your crawl (e.g. no. of URLs in Crawldb) works better by using the 'readdb’ command in the bin/nutch script, e.g. getting the crawldb statistics: *bin/nutch readdb <crawlDir> -stats*